PATENT
450117-04828

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICATION FOR LETTERS PATENT

TITLE: METHOD FOR PROCESSING SPEECH USING ABSOLUTE LOUDNESS

INVENTORS: Thomas KEMP, Ralf KOMPE, Raquel TATO

William S. Frommer
Registration No. 25,506
FROMMER LAWRENCE & HAUG LLP
745 Fifth Avenue
New York, New York 10151
Tel. (212) 588-0800

## Description

The invention relates to a method for processing speech, in particular to a method for emotion recognition and speaker identification.

In many systems with man-machine interfaces (MMI) it is desirable to integrate as much information as possible that can be derived from the various communication channels used by humans. In particular, it is often useful to include emotional information that describe the emotions of a user of a system, i.e. for example if the user is angry, happy, or sad. This emotional information may be derived from a speech signal of the user and can then be used e.g. to generate a respective response of the system. An example for a system, where emotional information can be useful, is an automatic teller machine (ATM) which is speech operated. If the user gets annoyed by the system, because the system has e.g. asked the user to repeat an order several times, he may get impatient. This emotional state may be detected by the system and thus the system's input mode may switch from speech to graphic/haptic input via a touch screen.

Another important point of today's MMI systems is the identification of speakers. In many systems it is important to know who is interacting with the system. For example, several people may share a car and certain parameters of the system may be set dependent on the current driver. It is therefore necessary that the driver be identified, which is commonly achieved by a speaker identification routine within the MMI system.

It is an object underlying the invention to provide a method for processing speech, in particular for emotion recognition and/or speaker identification.

To achieve this object, the invention provides a method according to claim 1. In addition, the invention provides a speech processing system, a computer program product, and a computer readable storage medium as defined in claims 9, 10, and 11, respectively. Further features and preferred embodiments are respectively defined in respective subclaims and/or in the following description.

According to the invention, a method for processing speech comprises the steps of receiving a speech input of a speaker, generating speech parameters from said speech input, determining parameters describing an absolute loudness of said speech input, and evaluating said speech input and/or said speech parameters using said parameters describing the absolute loudness.

This means, absolute loudness is used during evaluation of said speech input in addition to other parameters typically used in a classifier (e.g. a classifier for determining an emotional state of said speaker), such as prosodic features or voice quality features. Quality features of speech, i.e. auditory features arise from variation in the source signal and vocal tract properties, which are very speaker dependent.

Preferably, the step of evaluation comprises a step of emotion recognition and/or a step of speaker identification. The use of absolute loudness as a parameter for emotion recognition and speaker identification is a key feature of the invention. The rate of successful emotion recognition and the rate of successful speaker identification improved significantly using absolute loudness as an additional input parameter for the respective regognition systems.

Advantageously, a microphone array comprising a plurality of microphones, i.e. at least two microphones, is used for determining said parameters describing the absolute loudness. With a microphone array the distance of the speaker from the microphone array can be determined by standard algorithms and the loudness can be normalized by the distance.

This is done by estimating a time difference between microphones using correlation techniques.

Further, a location and/or distance of the speaker is determined and the absolute loudness is determined using standard algorithms for auditory and/or binaural processing. Thereby, an artificial head or similar shape with two microphones mounted at ear position is used. Processing in the ear is simulated, i.e. time delay and amplitude difference information between two "ears" is estimated and used to determine exactly the speakers position..

Said absolute loudness is preferably computed by normalizing the measured loudness at the microphones (signal gain) or the energy by said distance. Preferably this is done by multiplication, i.e. Distance times Energy.

Said distance is thereby determined using standard algorithms for speaker localization. The normalization by the distance is a key feature of the invention, because the normalization transforms the measured loudness into the absolute loudness. By normalizing the loudness by the distance the determined absolute loudness becomes indedepent of the distance of the speaker from the microphone.

In prior art emotion recognition systems and speaker identification systems loudness could not be used because a speaker speaking with the same loudness appeared to speak with a different loudness depending on his distance to the microphone.

A speech processing system according to the invention is capable of performing or realizing the inventive method for recognizing speech and/or the steps thereof.

A computer program product according to the invention, comprises computer program means adapted to perform and/or to realize the inventive method of recognizing speech and/or the steps thereof, when it is executed on a computer, a digital signal processing means, and/or the like.

A computer readable storage medium according to the invention comprises the inventive computer program product.

The invention and advantageous details thereof will be explained by way of an exemplary embodiment thereof in the following with reference to the accompanying drawings, in which

**Fig. 1** shows a flowchart illustrating the inventive steps; and

**Fig. 2** shows an example of a localized speaker according to the invention.

In Fig. 1 the speech input SI of a speaker S, in the following also refered to as user of the system, is received by a microphone array MA. Then, speech parameters SP are derived from the received speech input SI. The speech parameters can be any kind of acoustic features, derived from the spectrum and/or time series, e.g. voice quality features, prosodic features,

In a distance computing step CD the distance D of the speaker S from the microphone array MA is determined, i.e. the speaker is localized. Thereby, the time difference TD (also referred to as time delay) between microphones is estimated, using correlation techniques.

The distance D is further used in a loudness computing step CL, wherein the absolute loudness L is determined, which is measured in units of dbA. The absolute loudness L is determined using the signal energy, i.e. the absolute loudness is the energy normalized by the distance D.

Thereby, signal energy is measured in a window, e.g. by

$$E = \sqrt{\sum_{n=1}^{N} s_n^2} \ .$$

where $s_n$ is the digistized speech signal. Many alternative formulars exist. In a similar way the signal energy E can be computed from the spectrum. In that case frequency based weighting according to ear sensitivity in different frequency bands can be applied. Since energy decreases proportional to 1/D, with D being the distance D between the speaker and the microphone (see Fig. 2), absolute energy or loudness can be computed as D*E, i.e. by multiplying the distance D and the Energy E.

The absolute loudness L is now used in an evaluation step EV. This evaluation step EV may comprise a speaker identification and/or emotion recognition.

Besides the absolute loudness L, standard features ESF are used in the evaluation step EV. These standard features ESF are extracted in parallel to the distance computing step CD, and the loudness computing step CL in a standard feature extracting step SFES. In this standard feature extracting step SFES the received speech parameters SP from the microphone array MA are processed.

In Fig. 2 a speaker S is shown. The speaker has a certain distance D from the microphone array MA. As mentioned above, the distance D is determined in the distance computing step CD of Fig. 1 and is used for determining the absolute loudness L. Thereby, the time difference of the signals arriving at the different microphones of the microphone array is determined using correlation techniques.

It should be noted that loudness could not be used in prior art systems for emotion recognition and/or speaker identification because in prior art systems only one microphone is used. If only one microphone is used, the loudness depends on the distance of the speaker to the microphone. Moreover, in prior art systems, the speech signal is normalized to eliminate any "disturbing" variance of loudness. This fact further prevents the use of loudness for emotion recognition and/or speaker identification.

With the invention, the absolute loudness can now be determined and be used for emotion recognition and/or speaker identification. In this context it is assumed that absolute loudness can be important for emotion recognition and also is char-

acteristic for speakers and thus carries valuable information for speaker identification.

**Reference Symbols**

| | |
|---|---|
| **CD** | distance computing step |
| **CL** | loudness computing step |
| **D** | distance |
| **ESF** | extracted standard features |
| **EV** | evaluation step |
| **L** | absolute loudness |
| **MA** | microphone array |
| **S** | speaker |
| **SFES** | standard feature extracting step |
| **SI** | speech input |
| **SP** | speech parameters |
| **TD** | time difference |